Transforming the Way
Researchers Share Data

LESSONS FROM THE PEDIATRIC CANCER DATA COMMONS
Slides available at http://sam.am/CAC2-2020
Learning points

- **Data standards**: Data standards are a fundamental building block in harnessing the potential of big data.

- **Pediatric-specific challenges**: Pediatric cancer presents unique challenges and is lagging behind, in large part because of the lack of pediatric-specific data standards - particularly clinical data standards.

- **Consensus**: Defining and maintaining data standards require deliberation, debate, and consensus from across the pediatric cancer spectrum and medical specialties.

- **Ecosystems and collaboration**: Harnessing the potential of big data requires an ecosystem of federated systems that can work together to advance the cause.
A Perfect Storm: Rare diseases, siloed data, and manual processes
Pediatric cancer is a rare disease

18 million new cases of cancer worldwide every year

224,000 (1%) are children

Sources: Globocan 2018 (adult), Ward 2019 (pediatric)
Ultimately, this is what we’re trying to fix
Current state: Lack of standardization

Ways that age is expressed in the Gene Expression Omnibus (GEO)

<table>
<thead>
<tr>
<th>age</th>
<th>age [y]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>age [year]</td>
</tr>
<tr>
<td>AGE</td>
<td>age in years</td>
</tr>
<tr>
<td><code>Age</code></td>
<td>age of patient</td>
</tr>
<tr>
<td>age (after birth)</td>
<td>Age of patient</td>
</tr>
<tr>
<td>age (in years)</td>
<td>age of subjects</td>
</tr>
<tr>
<td>age (y)</td>
<td>age(years)</td>
</tr>
<tr>
<td>age (year)</td>
<td>Age(years)</td>
</tr>
<tr>
<td>age (years)</td>
<td>Age(yrs.)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>Age, year</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>age, years</td>
</tr>
<tr>
<td>age (yr)</td>
<td>age, yrs</td>
</tr>
<tr>
<td>age (yr-old)</td>
<td>age.year</td>
</tr>
<tr>
<td>age (yrs)</td>
<td>age_years</td>
</tr>
</tbody>
</table>

Adapted from Mark A. Musen, M.D., Ph.D.
What is a data commons?

Clinical data
Genomic data
Biospecimen data
Analytics tools

Researcher puts patient data into commons

Researcher puts lab data into commons

Researcher puts genomics data into commons

Researcher can use and analyze all data right in the commons
What kinds of research can a commons enable?

- Correlating **biomarkers** with clinical outcomes across trials
- Understanding impact of **dose modifications** across trials
- Performing patterns of failure analyses
- Examining **toxicity** prognosticators

- Validating consensus **staging definitions** across trials
- Validating **prognostic scoring systems**
- Enhancing **risk stratification**

- **Prognosis** of rare subgroups
- **Age-related differences** in therapeutic response
- **Survivorship** studies
- **Disparities** analyses

mskcc.org

Macmillan Cancer Support

mskcc.org
Essential elements for building a commons

**CONSORTIUM**
building trust between groups

**DATA GOVERNANCE**
publication policy, approving data use

**DATA TRANSFORMATION AND AGGREGATION**
statisticians and data scientists

**SCIENTIFIC GOALS**
why build it? what data to include?

**DATA DICTIONARY**
everyone speaking the same language

**FUNDING**
building and sustaining the commons
Our first commons (2015): Neuroblastoma

INTERNATIONAL NEUROBLASTOMA RISK GROUP (INRG), 2004

<table>
<thead>
<tr>
<th>Group</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG</td>
<td>4235</td>
</tr>
<tr>
<td>Germany</td>
<td>1938</td>
</tr>
<tr>
<td>Japan</td>
<td>470</td>
</tr>
<tr>
<td>SIOPEN</td>
<td>936</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8800</strong></td>
</tr>
</tbody>
</table>

The good news: 8800 patients

The bad news: data in Excel
Neuroblastoma data commons - Cohort discovery

Cohort Search

Publicly available at http://inrgdb.org/
Imaging data
Visualization and analytic tools

Visualization

Survival analysis parameters:
- Survival time limit (in years): 0 - 10
- Event Free/Overall Survival: Overall Survival
- Select a factor variable: Ethnicity
- Select a stratification variable: - Please select -

Cohort selection criteria:
- INSS/Stage: Stage 4

Results
Consensus-based decision making and data sharing
International Soft Tissue Sarcoma Consortium (INSTRuCT)

Copenhagen
May 2017

Chicago
October 2017

Amsterdam
March 2018

Tübingen
September 2018

Utrecht
December 2018

Prague
May 2019

Transforming the Way Researchers Share Data
October 21, 2020 - http://sam.am/CAC2-2020

@PedsDataCommons
http://commons.uchicago.edu
Data dictionary development
Data dictionary development
Balloting a consensus data dictionary

Disease consortia:
- COG
- MMT
- CWS
- EpSSG
- AIEOP

Case report forms → Multiple data dictionaries → Common data dictionary → Harmonized case report forms

Harmonize completed trials using new dictionary

New data flow easily into the commons

Data standards resources are updated

caDSR NCIt

Data commons
# Data standards example

## Neuroblastoma stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 = Stage 1</td>
<td>The tumor is confined to the original site of growth; no evidence of metastasis</td>
<td>C85417</td>
</tr>
<tr>
<td>2 = Stage 2a</td>
<td>The tumor is unilateral and the resection is grossly incomplete</td>
<td>C85418</td>
</tr>
<tr>
<td>3 = Stage 2b</td>
<td>The tumor is unilateral and the resection is complete or no resection</td>
<td>C85419</td>
</tr>
<tr>
<td>4 = Stage 3</td>
<td>The tumor extends across the midline and the regional lymph nodes are involved</td>
<td>C85420</td>
</tr>
<tr>
<td>5 = Stage 4</td>
<td>Tumor spread to distant lymph nodes, bone marrow, blood, or distant organs</td>
<td>C85421</td>
</tr>
<tr>
<td>6 = Stage 4s</td>
<td>Patients are less than one year old with localized primary tumor</td>
<td>C85422</td>
</tr>
<tr>
<td>9 = Unknown</td>
<td>Not known, not observed, not recorded, or refused</td>
<td>C17998</td>
</tr>
</tbody>
</table>
Pediatric Cancer Data Commons
Expansion
Guiding principles governing the PCDC

1. **OUR GOAL**
   is to lift barriers and connect researchers to data

2. **STAKEHOLDER APPROVAL**
   for data release from any disease commons

3. **CONTRIBUTOR APPROVAL**
   for data release from the original contributing group

4. **REPRESENTATION**
   on the Executive Committee from every disease group

5. **RECOGNIZE REGIONAL DIFFERENCES**
   Privacy and data sharing laws vary by region
An integrated pediatric cancer data commons

Clinical Data

- ALL
- AML INTERACT
- CNS
- DIPG IDIPGR
- EWS/OS HIBISCus
- GCT MaGIc
- HL NODAL
- NBL INRG
- STS INStruCT

Liver tumors
Retinoblastoma

PCDC Data Platform

Imaging
Biobank data
Genomic data

Other commons

NOTE: The PCDC can house the clinical data or link to clinical data housed in other databases.
### PCDC progress to date

<table>
<thead>
<tr>
<th>Disease</th>
<th>Stakeholders Engaged</th>
<th>Data Dictionary Established</th>
<th>Data Contributors Committed</th>
<th>Consortium MOU Signed</th>
<th>Cases in Commons</th>
<th>Analyses in Progress</th>
<th>Papers Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Lymphoblastic Leukemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute Myeloid Leukemia</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bone Tumors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNS Embryonal Tumors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diffuse Intrinsic Pontine Glioma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germ Cell Tumors</td>
<td></td>
<td>V1</td>
<td>8</td>
<td>MaGIC</td>
<td></td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td>Hodgkin Lymphoma</td>
<td></td>
<td>V1</td>
<td>2</td>
<td>NODAL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroblastoma</td>
<td></td>
<td></td>
<td>4</td>
<td>INRG</td>
<td>22K</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Soft Tissue Sarcoma</td>
<td></td>
<td>V2</td>
<td>5</td>
<td>INSTRuCT</td>
<td>4.6K</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Integration with national efforts
NCI Cancer Research Data Commons ecosystem

Credit: Allen Dearry
Center for Cancer Data Harmonization (CCDH)

- **Facilitate** retrospective and prospective semantic harmonization of data across nodes of the CRDC
- **Coordinate** the community to ensure implementation of standards that will facilitate interoperability of heterogeneous data types and CRDC resources
- **Find agreement** across the communities built around CRDC
  - match and extend data models
  - annotation, harmonization
  - quality assurance
Closing Global Gaps in Pediatric Cancer Survival
Research to date does not benefit all children equally

- 80% of children with cancer diagnoses live in low- and middle-income countries
- Research to date does not capture diversity of the global population

Survival rate disparity

Five-year age-standardized net survival (%) for lymphoid leukemia

Opportunities for the PCDC to close global gaps

- Continue to promote and incorporate consensus data standards development globally

Recent impact:
- INRG - recent publication
- MaGIC - working with researchers

---

**Pediatric Blood & Cancer**

ONCOLOGY: RESEARCH ARTICLE

The prognostic strength of serum LDH and serum ferritin in children with neuroblastoma: A report from the International Neuroblastoma Risk Group (INRG) project

Veronica Moroz, David Machin, Barbara Hero, Ruth Ladenstein, Frank Berthold, Paige Kao, Yaa Obeng, Andrew D.J. Pearson, Susan L. Cohn, Wendy B. London

First published: 30 May 2020 | [https://doi.org/10.1002/pbc.28359](https://doi.org/10.1002/pbc.28359)

**Conclusions**

LDH and ferritin are strongly prognostic in NB, overall and within high-risk NB patients treated post-2009 with modern therapy. LDH and ferritin show promise for (a) identifying ultra-high-risk; (b) refining risk stratification; and (c) clinical utility in low/middle-income countries. Routine collection of LDH and ferritin should be reinitiated for evolving NB risk stratification.
Targeting Survivorship
Survivorship

- A barrier to studying survivorship and long-term outcomes is access to longitudinal data
- PCDC is working with various stakeholders to create better ways to collect data on survivors
- Potential novel data collection methods include crowdsourcing
- A long-term outcomes / survivorship data commons is an essential component of our pediatric cancer strategy
Take home points

• Studying pediatric cancer requires **collaboration and sharing**
• Data sharing needs to be built on a foundation of **trust and consensus**
• Connecting disparate data types and sources enriches research
• **Consensus data standards** are critical for the success of national and international data ecosystems - allowing aggregation across trials and diseases
• **Early adoption of data standards** and consideration for the lifecycle of the data is critical to accelerating progress and discovery
The Pediatric Cancer Data Commons team

Suzi
Regulatory

Luca
Programmer

Sam
PI & Pediatric Oncologist

Monica
Director of Operations

Nicole
Data Standards

Anoop
Analytics

Jian
Programmer

Kat
Project Manager

Sam
PI & Pediatric Oncologist

Monica
Director of Operations

Caitlin
Communications

Maura
Data Standards

Not pictured: Sarah (Data Standards), Bobae (Front End Developer), Tom (Full Stack Developer), Shazia (Technical PM)
We gratefully acknowledge our funders

<table>
<thead>
<tr>
<th>Sammy's Superheroes</th>
<th>Leukemia &amp; Lymphoma Society</th>
<th>St. Baldrick's Foundation</th>
<th>Rally! Foundation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Neuroblastoma Children's Cancer Society</td>
<td>William Guy Forbeck Research Foundation</td>
<td>Andrew Tobin</td>
<td>Brightside</td>
</tr>
<tr>
<td>Kick Cancer</td>
<td>Children's Neuroblastoma Cancer Foundation</td>
<td>The Super Jake Foundation</td>
<td>King Baudouin Foundation</td>
</tr>
<tr>
<td>Alex's Lemonade Stand</td>
<td>Children's Research Foundation</td>
<td>Little Heroes</td>
<td>Mullin Fund</td>
</tr>
<tr>
<td>At the Forefront of Medicine</td>
<td>Children's Comprehensive Cancer Center</td>
<td>A gift made in memory of Payton O'Brien</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>UChicago Medicine</td>
<td>The Brumfield Family</td>
<td>UChicago Medicine Comprehensive Cancer Center</td>
<td></td>
</tr>
</tbody>
</table>